Dynamic Operator Switching with Walsh–Hadamard Weight Reuse

Overview

A standard ReLU network applies a learned linear transform and then discards inactive channels:

$$y = AWx$$

where A is a diagonal 0/1 activation mask determined by the ReLU decisions.

This creates sparse nonlinear behavior, but it also destroys information:

- inactive neurons become exactly zero,
- gradients disappear through inactive channels,
- and the effective Jacobian can rapidly lose rank.

An alternative viewpoint is to treat the activation mask not as a mechanism for *deleting computation*, but as a mechanism for *switching between computational operators*.

This leads naturally to constructions such as:

$$AWH + (I-A)HW$$

$$AWH - (I-A)HW$$

and

$$AWH + (I-A)HWH$$

where:

- W is a learned weight matrix,

- H is the Walsh–Hadamard transform,
- and A is a diagonal state-dependent gating matrix.

The central idea is simple but potentially powerful:

> inactive channels do not become zero — they switch to an alternative transform.

This changes the role of nonlinear gating completely.

Instead of:

- “compute or die,”

the network becomes:

- “compute differently depending on state.”

That may significantly improve:

- information preservation,
- parameter efficiency,
- activation diversity,
- and spectral richness.

---

1. Background

ReLU as Operator Deletion

Standard ReLU networks can be written as:

$$y = AWx$$

with:

$$A = \mathrm{diag}(a_i),\ a_i \in \{0,1\}$$

Inactive coordinates are removed entirely.

Geometrically:

- ReLU selects half-spaces,
- but also deletes dimensions locally.

This deletion is one of the main causes of:

- dead neurons,
- irreversible information loss,
- and rank collapse.

Walsh–Hadamard Transform

The Walsh–Hadamard transform (H) is:

- orthogonal,
- self-inverse,
- fast O(nlog n),
- and composed entirely of (±1).

For normalized H:

$H^T H = I$

And:

$H^{-1} = H$

It acts as a global sign-structured coordinate remixing operation.

Because:

$WH \neq HW$

For generic W, composing W with H in different orders creates distinct operators from the same learned parameters.

---

2. Option 1: AWH + (I-A)HW

Definition

M_1 = AWH + (I-A)HW

Coordinate-wise:

$$y_i = \begin{cases} (WHx)_i & \text{if active} \\ (HWx)_i & \text{if inactive} \end{cases}$$

Interpretation

This is the most conservative and probably most practical variant.

The activation mask no longer determines:

- whether computation occurs,

but instead determines:

- which operator is used.

Inactive channels remain computationally alive.

Key Property

Rewrite:

M_1 = HW + A(WH-HW)

This reveals:

- HW is the baseline operator,
- activation applies local corrections toward WH.

The layer therefore behaves like:

- a dynamically interpolated operator field.

Why It Is Interesting

Unlike standard ReLU:

- no coordinates collapse to zero,
- the effective Jacobian may retain substantially better conditioning,
- and all channels remain involved in information propagation.

This may improve:

- gradient flow,
- signal preservation,
- and expressive efficiency per parameter.

Conceptually, this begins to resemble:

- conditional computation,
- dynamic routing,
- or lightweight mixture-of-experts systems.

---

3. Option 2: AWH - (I-A)HW

Definition

M_2 = AWH - (I-A)HW

$$y_i = \begin{cases} (WHx)_i & \text{if active} \\ -(HWx)_i & \text{if inactive} \end{cases}$$

Interpretation

This variant is substantially more aggressive.

Inactive channels do not merely switch operators:
they switch operator *and polarity*.

Structural Rewrite

M_2 = -HW + A(WH+HW)

Now inactive regions perform signed reflections of the alternate operator.

This introduces:

- antisymmetric structure,
- sign-sensitive routing,
- and potentially oscillatory dynamics.

Why It Is Interesting

This variant may generate:

- richer spectral behavior,
- stronger polarity preservation,
- and more complex piecewise-linear partitions.

It starts resembling:

- wave interference systems,
- signed transport networks,

- or reversible dynamical systems,

more than conventional sparse feature detectors.

Potential Risks

The same mechanisms that increase expressivity may also create:

- optimization instability,
- cancellation effects,
- or chaotic depth behavior.

Residual pathways may be important for stabilization.

---

4. Option 3: AWH + (I-A)HWH

Definition

M_3 = AWH + (I-A)HWH

Coordinate-wise:

$$y_i = \begin{cases} (WHx)_i & \text{if active} \\ (HWHx)_i & \text{if inactive} \end{cases}$$

Spectral Interpretation

This is perhaps the most conceptually elegant construction.

Notice:

HWH

is W expressed in the Hadamard basis.

So the network dynamically switches between:

- one operator acting in the original coordinate system,
- and another acting in a transformed spectral coordinate system.

Interpretation

Active channels use:

- standard mixed-domain processing.

Inactive channels instead route through:

- a spectrally conjugated version of the weights.

This creates something like:

- adaptive basis selection,
- or dynamic spectral routing.

Why It Is Interesting

The layer can be interpreted as dynamically choosing between:

- spatial-domain computation,
- and Hadamard-domain computation.

This is a much richer mechanism than ordinary ReLU gating.

The same learned parameters are reused in:

- multiple coordinate systems,
- multiple sign structures,
- and multiple activation geometries.

---

5. Shared Properties

All three constructions share several important characteristics.

5.1 Operator Switching Instead of Zeroing

This is the key conceptual shift.

Standard ReLU:

- deletes inactive channels.

These constructions:

- redirect inactive channels into alternative transforms.

That may significantly reduce information destruction.

5.2 More Nonlinear Structure Per Parameter

The same learned matrix W participates in multiple effective operators:

WH, HW, HWH

Because the transforms do not commute, each branch creates different activation geometries.

This increases:

- nonlinear partition richness,
- activation diversity,
- and functional complexity per parameter.

5.3 Orthogonal Mixing

The Walsh–Hadamard transform preserves norm:

$$||Hx|| = ||x||$$

So the mixing itself does not inherently destroy information.

Most information loss therefore comes from gating structure rather than the transforms.

5.4 Exponentially Many Effective Operators

For an n-dimensional activation mask, there are potentially:

$$2^n$$

distinct operator configurations.

The network therefore becomes a dynamic operator-selection system rather than a fixed affine transform followed by thresholding.

---

6. Which Variant Is Most Promising?

AWH + (I-A)HW

Most likely:

- easiest to optimize,
- best conditioned,
- and most practical.

A strong candidate for initial experiments.

AWH - (I-A)HW

Potentially:

- the most expressive,
- but also the least stable.

Interesting for studying:

- antisymmetric dynamics,
- signed transport,
- or reversible behavior.

AWH + (I-A)HWH

Probably the most intellectually interesting.

It introduces:

- dynamic basis switching,
- spectral-domain routing,
- and adaptive coordinate systems.

This may produce surprisingly rich behavior despite minimal parameter overhead.

---

7. Experimental Directions

Several measurements would be particularly informative:

- singular value propagation through depth,
- Jacobian conditioning,
- activation entropy,
- dead-channel statistics,
- branch correlation,
- gradient norm preservation,
- and expressivity per parameter.

Especially interesting would be comparisons against:

- standard ReLU MLPs,
- orthogonal networks,
- gated residual systems,
- and structured-matrix architectures.

8. Final Perspective

These constructions suggest a different way of thinking about neural nonlinearities.

Instead of:

- “neurons either activate or disappear,”

the network becomes:

- a dynamic system that reroutes computation through different operators depending on state.

That shifts the role of gating from:

- sparse deletion,

to:

- adaptive operator selection.

The Walsh–Hadamard transform is particularly attractive for this because it is:

- orthogonal,
- cheap,
- sign-structured,
- and highly mixing.

As a result, these architectures may occupy an interesting middle ground between:

- conventional ReLU networks,
- reversible systems,
- spectral methods,
- and conditional computation architectures.

They appear simple enough to implement immediately, but structurally unusual enough that their behavior may differ substantially from standard deep networks.